CMPS142-Spring 2018
Homework 2

Handed out: May 8, 2018
Due: May 17, 2018 at 1:30 PM

- You are allowed to solve this homework in groups of 2 or 3. Collaborating with any one not enrolled in the class, (except the course staff), or taking help from any online resources for the homework problems is strictly forbidden.
- One (and only one) member of the group has to submit the homework using his/her account on canvas. All group members will get points for that submission.
- How to submit your solutions: Your group's solution to each problem must be typed up separately (in at least an 11-point font) and submitted in the appropriate 'Problem' box on the Canvas website as a PDF file. **This means that if the homework has N problems, you will submit N separate pdf files on Canvas, one for each problem per group!** For example, submit the pdf that contains your group's solution to the first problem in the box titled 'Problem 1'.
- **Each pdf file** should clearly mention the names, email addresses and student ids of all group members. If you forget a group member's name, they will not get points for that problem.
- It is your responsibility to ensure that the files that you submit are not corrupted in any way. After you submit a file, please download it and verify that there aren't any issues. Any related requests will not be entertained after the due date.
- You are very strongly encouraged, but not required, to format your solutions in LATEX. You can use other softwares but handwritten solutions are not acceptable.
- Please try to keep the solution brief and clear.
- The homework is due at 1:30 PM on the due date. There is a 10% penalty (10 points) for each late day, upto 3 days. After that you will not get any points for this homework. Note that your submission will be considered late even if you are late for one (or more) of the problems.
- The Computer Science Department of UCSC has a zero tolerance policy for any incident of academic dishonesty. If cheating occurs, consequences within the context of the course may range from getting zero on a particular homework, to failing the course. In addition, every case of academic dishonesty will be referred to the student's college Provost, who sets in motion an official disciplinary process. Cheating in any part of the course may lead to failing the course and suspension or dismissal from the university.

Problem 1: Support Vector Machines [30 Points]

1. Consider a binary classification dataset with two features, x_1 and x_2, shown in the table below. + represents the positive class and - represents the negative class.

   We now want to train a Hard-margin SVM classifier, which learns decision boundaries that lead to the largest possible margin from the two classes. In general, this can be a hard optimization problem. But for this small dataset, we can solve this by manual inspection.

x_1	x_2	Label
1	1	+
2	2	+
0	4	+
2	4	-
4	2	-
5	4	-

(a) [3 points] Plot the dataset. Are the two classes linearly separable? Which of the points are support-vectors?

(b) [4 points] What will be the optimal weight vector w and bias b for this dataset? Draw the corresponding decision boundary in the plot.

(c) [3 points] What is the size of the maximum margin?

(d) [3 points] What happens to the size of the maximum margin if we remove one of the support vectors from the dataset? Does it increase, decrease, or remain the same? Provide a brief justification why.

(e) [3 points] What happens to the decision boundary if a point that is not a support vector is removed from the dataset? Briefly explain why.

2. **Soft-margin SVMs:** The hard-margin SVM model can only learn a solution for cases when data are separable. In real data applications, this is usually unrealistic. To overcome this, the soft-margin version of SVMs (discussed in class) relaxes this assumption by introducing a slack variable ξ_i for each point (x_i, y_i) in the data. This leads to the following modified optimization problem:

$$\min_{w,b,\xi_i} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i$$

$$s.t. \ \forall\, i, \ \ y_i(w^T x_i + b) \geq 1 - \xi_i \quad \text{and } \xi_i \geq 0$$

(a) [3 points] Consider the points which are correctly classified and are not support-vectors. What is the value of ξ_i for such points? Justify your answer.

(b) [6 points] Prove that using this formulation:

$$\text{\#of mistakes in the training set} \leq \sum_i \xi_i$$

(c) [4 points] For soft-margin SVMs, if we train models with two values of C_1 and C_2, such that $C_1 < C_2$, in which case would you expect a larger number of mistakes on the training set, and why?

(d) [4 points] Can we think of the soft-margin SVM case as reducing to the hard-margin SVM for some value of C? Justify your answer.

Problem 2: Kernels and SVMs [10 points]

In this question, we will study feature representations with kernels. Consider the 1-dimensional dataset in Figure 1. It consists of 5 points, of which three are labeled positive and two are labeled negative.

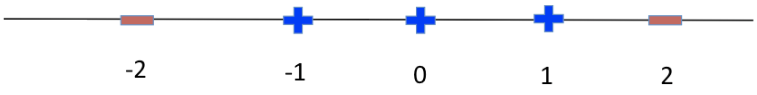

Figure 1:

1. [3 points] Is the data linearly separable? If not, find a feature map $\phi : \mathbb{R}^1 \rightarrow \mathbb{R}^2$, which maps points in the original 1-d representation *to a 2-dimensional feature representation so that the data becomes separable.* Plot the transformed dataset in 2-D.

2. [2 points] Intuitively, what will be the decision boundary for a hard-margin SVM in this 2-dimensional feature space? Draw this decision boundary in your plot.
3. [2 points] What will this decision boundary look like in the original 1-D space?
4. [3 points] You just trained a kernel classifier. For the feature map ϕ that you devised above, what is the corresponding kernel function $K(x, y)$?

Problem 3: Experiments with Weka [60 Points]

In this experiment, we will study the properties of some classifiers we have learned about in the class using Weka (3.8.2). In particular, you will be experimenting with KNNs, Naive Bayes, Logistic Regression, and SVMs. Some questions include the performance figures for Decision Trees to help you verify your experimental settings. This is only included to help you verify your settings and the question is NOT about Decision Trees. All experiments (except the last one) have to be done for the dataset provided with this homework (cmps142S18_HW2.arff). The questions are about building binary classifiers on the dataset in this file. There are 8 numeric attributes (or features) named *preg*, *plas*, *pres*, *skin*, *insu*, *mass*, *pedi*, and *age*. There is also 1 binary categorical variable named *class* that represents the class (or the label) for the classification problem. Note that there are fewer points for reporting the correct performance figure and more points for the associated subjective questions.

1. In this question you have to report the accuracies of various classifiers on the *training set*, and then answer questions about the performance. You can make Weka report training performance by selecting "Use training set" for "Test options". Use default settings for all classifiers.

   (a) [4 points] Complete Table 1 and include it in your report.

Algorithm Name	Where to find?	Train Accuracy
Decision Trees	`J48` under `trees`	84.1%
1NN	`IBK` under `Lazy`	
Naive Bayes	`Naive Bayes` under `bayes`	
Logistic Regression	`Logistic` under `functions`	
SVM	`SMO` under `functions`	

Table 1: Training Accuracies

   (b) [3 points] Which classifier has the highest accuracy? Why?

2. This question is about the Logistic Regression model whose performance you reported in Table 1. We learned in class that Logistic Regression yields a linear decision boundary of the form $\sum_i w_i x_i + b = 0$ where the subscript of i, represents the i^{th} feature (also called attribute), $w_i \in \mathbb{R}$ is the weight of the i^{th} feature, and b is the bias of the dataset. We also saw that the classification rule for Logistic Regression can be written as: `IF` $\sum_i w_i x_i + b > 0$ `THEN predicted_Class=`C .

   (a) [5 points] Write the equation of the decision boundary learned by Weka for Logistic Regression. (In the equation for the decision boundary provided above, you should replace x_i by the name of the i^{th} feature in the dataset, and w_i by the weight for that feature. Also replace b by the bias learned by Weka).

   (b) [1 point] What is the name of the most important feature for this dataset as learned by Logistic Regression? You can say that the most important feature is the one for which the absolute value of weight is highest. Also, bias is not considered a feature.

   (c) [3 points] Also write the classification rule for Logistic Regression as learned by Weka on this dataset. Use the general form of the rule provided above. Hint: you can print out the predictions for an instance by selecting a file format in `Output Predictions` which pops up when you click on `More options...` in the `Test options` window.

3. This question is similar to the previous one, except that it is about the SVM model whose performance you reported in Table 1. Like Logistic Regression, SVMs also learn a linear decision boundary of the form $\sum_i w_i x_i + b = 0$.

   (a) [2 points] Write the equation of the decision boundary learned by Weka for SVM.

   (b) [1 point] What is the name of the most important feature for this dataset as learned by SVM?

4. Now, you will report the 10-fold Cross Validation (CV) accuracy of each classifier. Notice that 10-folds is Weka's default setting for CV. It is also possible to change the number of folds using the text box in the GUI. However, this was only for your information and for this question and the rest of this assignment, you will work with 10-fold CV.

   (a) [4 points] Complete Table 2 and include it in your report.

Algorithm Name	10-fold CV Accuracy
Decision Trees	71.224%
1NN	
Naive Bayes	
Logistic Regression	
SVM	

Table 2: CV Accuracies

   (b) [3 points] For each classifier compare its 10-fold CV (Table 2) and Train (Table 1) accuracies. For which classifier do you see the biggest change? Why does this happen?

5. This question is about the 10-fold CV performance of Logistic Regression. Weka's implementation of Logistic Regression uses a `ridge` parameter $(-R)$ to *regularize* weights. In this question you will study the effect of this parameter on the learned model. You will try different values of R and for each value you will train a logistic regression model. You will then examine the learned model.

   (a) [2 points] Consider the following values of ridge: $\{0, 1, 10, 100, 1000\}$. How do the feature-weights of the learned model change with increasing values of ridge?

   (b) [EXTRA CREDIT 5 points] Draw a plot of R versus square of Euclidean norm of weights $(||w||_2^2)$. $||w||_2^2$ is defined as $\sum_i w_i^2$. Note that bias is usually not considered to be a part of the weight vector. Your plot should have $||w||_2^2$ on the y-axis and R on the x-axis. Consider the 5 value of R specified above.

   (c) Set $R = 10000$, set the 'test option' to 10-fold CV, and re-learn a Logistic Regression model.

      i. [4 points] Report the confusion matrix.

      ii. [3 points] You should see something strange about this confusion matrix (if you don't, compare it to the confusion matrix you obtained on 10-fold CV with default settings). What is strange about this matrix? How can you explain this strangeness?

6. This question is about the 10-fold CV performance of kNNs. In Weka, you can change the value of k in IBK using the parameter called `KNN` $(-K)$.

   (a) [4 points] Report the 10-fold CV performance of KNN using k{1,3,5}. What is the trend that you observe? Why do you see this trend?

   (b) Set $K = 1000$, set the `Test option` to 10-fold CV, and learn a KNN model.

      i. [4 points] Report the confusion matrix.

      ii. [3 points] You should see something strange about this confusion matrix. What is strange about this matrix? How can you explain this strangeness?

7. Now we will study the effect of duplicate (correlated) features on the performance of various classifiers. We are providing a file named 'cmps142S18_HW2_repeatedFeats.arff' with this homework. It is same as the 'cmps142S18_HW2.arff' file except that one of the features, `pres`, is duplicated 10 times. In this question, you will observe the 10-fold CV performance on this dataset and compare it to your results in Table 2.

   (a) [4 points] Complete Table 2 and include it in your report.

Algorithm Name	10-fold CV Accuracy
1NN	
Naive Bayes	
Logistic Regression	
SVM	

Table 3: CV Accuracies

(b) [6 points] Name the classifier(s) for which you see a significant change (of more than 1.0 point) in Accuracy as compared to Table 2. Explain this change(s).

(c) [4 points] Name the classifier(s) for which you DO NOT see a significant change (of more than 1.0 point) in Accuracy as compared to Table 2. Why weren't these classifier(s) affected by the duplicate features?